DOCUMENT RESUME

ED 036 830 CG 004 994

AUTHOR Nivette, James D.
TITLE A Rationale and Methodology for Designing Logical Evaluations for School Programs. Research Study Series, 1967--68.
INSTITUTION Los Angeles County Superintendent of Schools, Calif.
REPORT NO RR-5
PUB DATE 3 Jul 69
NOTE 23p.

EDRS PRICE EDRS Price MF-$0.25 HC-$1.25
DESCRIPTORS *Educational Objectives, *Evaluation Criteria, *Evaluation Methods, *Evaluation Techniques, Measurement Techniques, *Program Evaluation, Research Design, Standardized Tests, Statistical Analysis

ABSTRACT

Educational evaluation and design are discussed. Five essential steps in evaluation are: (1) define educational objectives in behavioral terms, (2) translate the objective into descriptions of behavior, (3) identify situations in which the designated behavior can be observed, (4) establish an interpretative device which can measure the desired growth and, (5) state conclusions regarding the extent to which the objectives were achieved. Two tables which present alternative methods for simple evaluation procedure designs are included. The development of objectives is also discussed. Objectives should: (1) describe what the student does, (2) describe conditions under which his performance can be observed, and (3) define the standards the student must meet. The evaluative process itself in the evaluation of school programs is considered next. Five evaluative designs are discussed including the use of control groups and standardized tests. Evaluative criteria dealing with interest might employ questionnaires, attendance records, case studies, etc.; a list of standardized tests, including personality, interest and achievement tests which are useful in evaluation. Lastly, a statistical refresher containing definitions of measurement terms and a discussion of the nature and purposes of statistics in relation to evaluation designs is given. (Author/EW)

Los Angeles County Superintendent of Schools
Division of Research and Pupil Personnel Services

Research Study Series
1967-68

Research Report Number 5

A RATIONALE AND METHODOLOGY FOR DESIGNING LOGICAL EVALUATIONS FOR SCHOOL PROGRAMS

Prepared by
Dr. James D. Nivette

A RATIONALE AND METHODOLOGY FOR DESIGNING LOGICAL EVALUATIONS FOR SCHOOL PROGRAMS

James D. Hivette

INTRODUCTION:

In response to the growing demand for logical, realistic educational evaluations (and since many of the educational projects being submitted each year to Washington are unacceptable because of weak evaluation designs) there is a need for educators to become aware of appropriate methods in evaluation designs. It is because of this need that a series of articles is being written as an expression of one opinion as to logical design and evaluation.

The first section is an introduction to evaluation - evaluation design. The second section deals with the development of objectives and with methods and procedures. The third section (probably the most important) deals with the evaluation design itself. Section four is a statistical refresher designed to help the individual gain insights into the nature and purposes of statistics in relation to evaluation designs. They may be each removed to comprise the total paper on evaluation.

The most important thing one can remember when designing an evaluation for a school project is that the evaluation must always aim to eliminate the effect of factors other than those that the project itself has on changes in the pupil as measured or observed during the course of the project. In other words, an evaluation should be designed so that factors other than the experiences gained in the projects themselves (error) are eliminated from consideration in the data.

Steps in Evaluation

In any evaluation there are five essential steps: Step 1: define the educational objectives logically in behavioral terms which are expected to be achieved through the experiences being evaluated. These objectives should reflect the most pressing needs of the students living in the area. Step 2: the educational objective should be translated into descriptions of behavior that will display that the objectives are achieved. These are the traditional behavioral objectives. Step 3: identify situations in which the presence or the absence of the designated behavior in relation to the objectives can be observed and recorded. Step 4: establish some type of interpretative device, standard or normative which can be used to measure the desired growth and is appropriate for the particular objectives being appraised. Step 5: state conclusions regarding the effectiveness of the program in terms of the extent to which the objectives were achieved as compared with the baseline data obtained at the beginning of the project.

The varied aspects of pupil growth along intellectual, personal, social, physical, and attitudinal lines suggest that a variety of assessment techniques may be needed to evaluate outcomes of a particular project or program. Every effort should be made to select or develop only those data gathering procedures that will provide information related to the specific objectives of the program. The following presents two alternative methods for simple evaluation procedure designs. In Table 1, one can see that the objectives or goals of the projects are clearly indicated. The objective is that youth will read at a level appropriate to his age, grade and intelligence. This objective is one that is measureable.

The methods designed to meet the goals are neatly listed. Following these methods are the means of evaluation. One can see that standardized test scores are being used on a pre-test post-test basis. This procedure is one of the best research designs. It is also evident from Table 1 that non-test sources of evaluation data are listed, which are useful and related in this instance.

Table 2, which is a somewhat more complex and yet not incomprehensible paradigm for evaluation, again shows how a chart can be used to make the understanding of evaluation designs much easier. In Table 2, one sees the learning outcomes which may be the objectives or sub-objectives of a project, the type of instrument or technique used to gather data in relation to the learning outcome, the responsibility for constructing that type of instrument, and the time when the instrument is used in the project. In that table it also is seen that there is a pre-post design. The subjects used in the study are identified for clarity. A last section is "remarks", which gives pros and cons to each evaluation technique.

Table 1

OBJECTIVES (Goals)	METHOD - CONTENT OUTLINE (To meet goals)	EVALUATION (List)	Instrument Procedures Technics
Example 1. To enable the educationally disadvantaged youth to read at a level appropriate to his age, grade, and intelligence.	1. Individualized instruction a. Specialized equipment b. Reading laboratory 2. Utilization of specialized staff a. Remedial reading specialist b. Psychologist c. Sociological services d. Medical services 3. Library facilities and materials a. Expanded reading list to encourage more reading	1.1 Standardized test scores a. Pre-post reading readiness b. Reading comprehension c. (List other standards utilized) 1.2 Non-test sources of evaluation data a. Teacher observation b. Daily teacher checklist c. Teacher-pupil conference d. Teacher evaluation checklist e. Pupil self-evaluation checklist Test Results for: a. Affective domain b. Social development c. Physical development 1.3 List proposed calendar or schedule of test and/or non-test sources of evaluation data	

Here, in simple step-by-step fashion, is the way in which an evaluation procedure might develop.

Table 2

Learning outcome	Type of instrument or technique	Constructed by	When used in project	With whom used	Remarks
The child has the ability to notice, name and distinguish things and events in his home environmen	Anecdotal records	Teacher	Beginning, middle stages, end (at least 3)	A sampling of 10 children	Sample might be selected on basis that the children are the most in need of improvement in this objective
The child is curious about, and takes a positive interest in learning to read	Rating scales	Guidance counselor or evaluation director	Beginning and end	All children	Can be quantified, made more useful if two or three raters can be brought in to make observations
The child notes essential details and generalizations in teacher explanation and pupil reports	Informal objective test	Teacher, collaborating with guidance counselor or evaluation director	Early in project; again at end	All children	Test may be all paper-and-pencil or a teacher report with paper-and-pencil response, depending upon groups
The child enjoys the experience of artistic expression in a variety of media	Anecdotal records	Teacher	As observed	All children	Teacher should make sure that there is at least one notation about each child, recording his apparent interest and his participation in at least one medium

A RATIONALE AND METHODOLOGY FOR DESIGNING LOGICAL EVALUATIONS FOR SCHOOL PROGRAMS (Cont'd.)

Section two in this series of articles deals with the development of objectives. It will also consider, briefly, methods and procedures. This article is directed toward a more complete understanding of the evaluation procedure.

OBJECTIVES

The most critical part of an evaluation design is precise objectives. One of the hardest evaluation jobs really begins at this time; that is, providing a clear statement of program objectives. The prime requirement of a statement of objectives is that they be clear and communicable to others. To reach this goal, objectives need to describe the behavior the student must perform at the end of the program, the conditions under which the behavior should appear, and the standards which the behaviors must meet. These three criteria clearly reflect a behavioral objective.

The first point, the achievement of the changes in behavior, is critical. The main idea about the objectives of the program is that they describe something children can do under observable conditions. Since any education program is composed of many things - curriculum, training aids, devices, etc., the performance of the student resulting from these program changes should be measured by various kinds of indexes written in performance tests. Clear and communicable objectives are necessary to insure that all these activities are contributing to the same goals. It is also imperative that three groups of people understand the objectives clearly - the students, the teachers, and the supervisors.

Objectives shouldn't be omitted merely because they are not measurable. When an objective can't be measured, there is often a temptation to exclude it from the statements of project objectives. This is a distortion of values. Objectives should be stated because they are important - measurability should not overshadow the objective itself.

Again, objectives should do the following: (1) describe what the student does, (2) describe the conditions under which his performances can be observed, and (3) define the standards the student must meet. Clearly, as was stated before, these three criteria go into the making of behavioral objectives. But there exists some confusion as to how to develop a behavioral objective. For one person, merely knowing what the criterion is for a behavioral objective is sufficient to enable him to pursue the development of such an objective. For others, a more detailed analysis is needed.

The first criterion for a behavioral objective is that it should describe what the student does. In this sense what we are talking about is a listing of expected results (as in Table 2). These would be the learning outcomes -- as clear a statement as is possible as to just what the student should be able to do when he finishes the program. One shouldn't try to measure such subjective things as artistic work or creative judgment. These are probably a waste of time. If one must, he can use indirect evidence of growth such as records of interest and participation. This is about as close as one can come to evaluating a subjective type of behavior.

Remember, unclear objectives lead to unclear evaluations. So when it comes to listing what a student does, it would be better to list several specific criteria that are measurable, than whole numbers of subjective behaviors.

When listing the expected results, remember the project may have more than one outcome; that is, a project aimed at a specific problem may produce a variety of different effects on other groups. Thus it is helpful when designing other methods of evaluation for the evaluator to try to anticipate what the secondary outcomes of the program will be and insure that these too are evaluated through some means. These secondary outcomes may give evidence as the fulfillment of a primary objective which is too subjective to handle objectively.

A second condition for a behavioral objective is that one is to describe contitions under which the performance of the student is to be observed. If any one direction can be made outstanding in this section, let it be that the evaluator must not use vague words. Many words and phrases very popular with writers of behavioral objectives are fairly vague and become such cliches that it is difficult to tell exactly what they mean and what behavior they imply. Words such as understands, believes, knows, appreciates, and accepts, are words which have become cliches in the development of behavioral objectives. These vague terms should be used only in summarizing statements, or in explicit descriptions of objectives. One should also try to avoid ambiguous terms; and instead, state as clearly as possible what the student should be able to do when he finishes the program. The use of such words as recite, list, match, distinguish between, or any other number of more specific terms which describe exactly what a person can do to show that he knows or understands, is most helpful. A sample of these kinds of words, organized in Bloom's Taxonomy (extremely useful in developing objectives) is given in Appendix.

In describing the conditions under which the performance is to be observed, one must also remember to answer the question of how the different kinds of conditions apply to a given objective and whether they make a difference. For this reason, it is important to know that the statement of a condition will restrict or broaden the amount of material the student has to learn. If the skill is performed differently under different circumstances; if the task is easier or more difficult under certain conditions than others - answers to these kinds of questions must be sought. Consider, for example, whether a student should be able to solve all the problems of a given type or only special kinds of problems. Do you want the student to be able to find the roots of any equation, or only the linear equations? -- measure any voltage, or only voltages between zero and 100? What we are saying here is that one should include a statement of the conditions affecting the task in the objectives whenever this will help to communicate to others any differences in the task created by special conditions.

The last criterion for behavioral objectives is the definition of the standard the student must meet. When we say "standard" we mean: must he attain 10 A's, or 5 A's?; must he attend school for 40 days to be successful, or just 20? -- a standard of performance the student must meet if he is to successfully complete the program. Two kinds of standards are needed: the first is the standard of accuracy (for example, what percent of problems must the student work correctly?); the other type of standard refers to the speed with which the student must perform. In many tasks time is of little consequence; in others, time may be critical.

Listed is an example of a completed behavioral objective. The section called listing expected results will be "A", the section describing conditions will be "B", and the setting of performance standards will be "C".

A - The student will write an essay in which he describes a contribution to the Constitution of the United States.

B - The five most important American political experts of the eighteenth century.

C - The essay will be between 10 and 15 pages in length; the student's selection of the five experts must be justified.

Additional Information

1 - He will be able to use any library resources he wishes; but is honor bound not to discuss his essay with others.

2 - The essay will be due one week from this date.

One final suggestion for meaningful objectives is that they should be easy to understand. This involves making outlines. The successful writer of objectives organizes his objectives in a hierarchy from the very general to the very specific. This is easy for everyone to understand. The very specific objectives may deal with the location at which the student is to perform the task, the particular equipment he may need and the type of measurement to be taken of his performance at that time. Remember always that the terminal objectives for one program may well be the sub-objectives when viewed from the standpoint of an entire course. This is easier to determine when the objectives are formed in an outline, or hierarchy with the more general objectives being described in detail by more specific objectives.

To many people this may seem trivial. Some may say, "Well, I could write ten thousand objectives for one English course". This may be true, but if one could even develop one hundred valid, measurable objectives with suitable evaluation criteria for the end of the semester, this would indeed be far ahead of the amount of evaluation that goes on now in English instruction. In this case, one may not have evaluated all aspects of English, but will have evaluated one hundred areas.

METHODS AND PROCEDURES

The subject of methods and procedures as it has been discussed many times by educators is one in which we are not all experts. There are many people in the districts, in the County Office, people in different consultant roles, and in the universities who can provide the expertise for developing methodology describing procedures. It is very important that these people be included in the development of the objectives for it is at this time that performance standards are set, the conditions for the observations of the behavior, which are the methods, are set; and the expected results are stated in the objectives. We may not know what the methods and procedures are per se, but through work as a team the methods and procedures can be integrated into the very heart of the objective.

A P P E N D I X

(Bloom's Areas)
(Cognitive and Affective)

A P P E N D I X

Schema for Synthesis of Major Contributions to Curriculum Construction

Bloom Taxonomy	Guilford Products	Guilford Operations	Taba Behavioral Objectives	Grouping
Knowledge	.Units .Classes	.Memory .Cognition	.To define. . . .To derive. . . .To identify. . . .To inquire. . . .To recall. . .	.Large group .Small group
Comprehension	.Relations .Systems	.Memory .Cognition	.To classify. . . .To debate. . . .To distinguish. . . .To recognize. . . .To translate. . .	.Small group
Application	.Relations .Systems	.Convergent .Divergent	.To acquire. . . .To develop. . . .To increase. . . .To organize. . .	.Independent study
Analysis	.Transformation .Implication	.Convergent .Divergent	.To analyze. . . .To compare. . . .To contrast. . . .To differentiate. . . .To experiment. . .	.Small group .Independent study
Synthesis	.Transformation .Implication	.Convergent .Divergent	.To combine. . . .To infer relation. . . .To modify. . . .To relate. . . .To synthesize. . .	.Small group .Large group
Evaluation	.Transformation .Implication	.Evaluation	.To deduce. . . .To document. . . .To evaluate. . . .To interpret. . . .To supplement. . .	.Independent study .Large group

SCHEMA FOR TERMINOLOGY BASED ON THE TAXONOMY OF EDUCATIONAL OBJECTIVES: AFFECTIVE DOMAIN

RECEIVING: 1.0	1.1 Awareness	To feel. . . . . To sense . . . . . To capture . . . . To experience. . .
	1.2 Willingness to Receive	To tend. . . . . To incline . . . . To tolerate. . . . To dispose . . . . To permit. . . . .
	1.3 Controlled or Selected Attention	To perceive. . . . To attend. . . . . To select. . . . . To favor . . . . . To prefer. . . . .
RESPONDING: 2.0	2.1 Acquiescence in Responding	To comply. . . . . To conform . . . . To acquiesce . . . To allow . . . . .
	2.2 Willingness to Respond	To cooperate . . . To volunteer . . . To offer . . . . . To contribute. . .
	2.3 Satisfaction in Response	To enjoy . . . . . To delight . . . . To profit. . . . . To gratify . . . . To satisfy . . . .
VALUING: 3.0	3.1 Acceptance of a Value	To believe . . . . To prize . . . . . To respect . . . . To esteem. . . . .
	3.2 Preference of a Value	To pursue. . . . . To seek. . . . . To want. . . . . To search. . . . . To elect . . . . .
	3.3 Commitment	To justify . . . . To convince. . . . To persuade. . . . To consign . . . .

ORGANIZATION: 4.0	4.1 Conceptualization of a Value	To examine. . . . . To clarify. . . . . To separate . . . . To isolate. . . . .
	4.2 Organization of a Value System	To create . . . . . To originate. . . . To integrate. . . . To interrelate. . . To systematize. . .
CHARACTERIZATION BY VALUE OR VALUE COMPLEX: 5.0	5.1 Generalized Set	To review . . . . . To revise . . . . . To re-examine . . . To predispose . . . To orient . . . . . To internalize. . .
	5.2 Characterization	To characterize . . To judge. . . . . . To resolve. . . . . To conclude . . . .

N.S. Metfessel, 1967

Section three is the consideration of the evaluative process itself in the evaluation of school programs.

EVALUATION

An evaluation section should stand alone. It is an entity in itself. It aims to measure the effect of the program while parceling out the other influences of the school program; so when the data is evaluated, one is evaluating what happened in the program and not extraneous influences. Some of these changes one will be able to evaluate simply; others will require sophisticated designs.

One of the trouble spots in an evaluation is that of an absolute standard with which comparison can be made. Thus, it is not always clear that any improvement in behavior is a direct result of any educational program or whether such things as halo affects, Hawthorne affects, or other types of error are reflected. Increased maturity of the student can even play a part in the development of a certain skill.

Five Simple Designs

There are different ways in which the research worker can use controls in evaluative designs so that measures of changes within the program can be adequately demonstrated. Five designs follow in this discussion.

The first of these is the use of standardized tests. The norms of these tests are generally not applicable to non-standard school populations such as the educationally disadvantaged. When using standardized tests, one should always check, through the use of a chi-square test, for the normalcy of distribution of local data in relation to the standardization sample. If it is significantly different, one should norm the test for the particular school and compare students in the school with themselves or their own local norms. (The obvious advantage to local standards is that the students are geographically comparable within themselves.) Standardized tests can also be used to establish baseline data. A pre-post research design is a tight design wherein a test is administered at the beginning of the program and again at the end of the program and changes are evaluated through the use of the t statistic, or one-way analysis of variance, depending upon the number of groups involved.

A second method of evaluation design is the use of control groups. While many people have urged the use of control groups as the best means of getting standard comparison data, it is very difficult to set up equivalent control groups in some projects because many of the available children may actually be in the program. It is ethically imperative to provide opportunities to all of the students, rather than set aside a comparison group as controls. Even if a control group were obtainable, it might be systematically different from the program group if the schools attended by the two groups were different. If both groups were kept in the same school, the control group might be influenced by the project group's program. As an example of this -- the teacher and her control group find out that they are the control group and thus become determined to make sure that they do as well as those in the experimental groups. The most important statistic in control group design is the t-ratio or Analysis of Variance and Covariance. In a control group design, whatever measurement one takes should be subjected to a test for significant mean difference at the beginning; for if there is a significant mean

difference in the performance of the control group and the experimental group at the beginning, then there certainly might be a significant difference at the end also. Even if there isn't, not very much is said by this. Through the use of different control groups and different experimental groups and different variables, the possibilities of using a variety of multivariate procedures open up.

A third type of evaluative design is the use of hypothetical standards. This is useful when a control group cannot be found. When a suitable control group cannot be made available, one or two alternatives might be considered: (1) a projected average score for the grade level one is measuring based on past years' scores, or (2) a projected average score of children at other grade levels. In both cases the actual scores obtained on the measures are compared with the projected average.

The fourth type of evaluation design is the use of change standards. If no comparative data is available, one can measure change resulting from a project by comparing scores earned when the project was completed with those earned at the beginning of the project. In effect, the program participants themselves are used as controls. If the project begins in midyear, measurements can be made three times -- at the beginning of the year, when the project begins at midyear, and at the close of the project year. When the program is carried on in several schools, one will be able to observe whether directional amount of change is similar from school to school. Then if the program is repeated in the same school, one will be able to compare changes in the two cycles -- between school and within school. With change standards, complex designs such as time series analysis can be used where complex statistical techniques must be employed to perform an analysis of the data. Time samples is another method which could be used, as well as multiple time series designs.

A fifth type of evaluation design is the use of single measurements. In this design one would compare a group with any other group, such as overachievers, underachievers, above average students, below average students, similar classes in previous years, classes of similar children receiving the usual program but not part of the project, test standardization samples, or classes in different kinds of projects. Comparisons have varied methods of reaching the same objectives and are most fruitful in yielding information upon which to base a guide for the project. For example, a remedial reading program that emphasizes skill training can be compared with one that places the emphasis on individual diagnosis of reading difficulties or on motivation problems. Through the use of different techniques cross-comparisons can be made thus enabling one to develop more realistic designs for future years.

A book which is extremely helpful in the design of research and evaluation is "The Encyclopaedia of Educational Research," in particular, Chapter 5 by Donald T. Campbell and Julian C. Stanley, entitled, "The Experimental and Quasi Experimental Designs for Research and Teaching." Although this chapter requires some knowledge of statistics, at least the preliminary sections can generally be understood by the person with only moderate versatility in this area.

Evaluation Criteria

We have talked about some evaluation designs; now we encounter a new problem. One designs an evaluation in order that the measurements taken may be statistically analyzed in such a way as to make the most out of the data gathered. But what is the data that is gathered? Such evaluation criteria can come from a number of different sources. A list such as this may be used: In the area of subject matter and skill achievement one could use (1) appropriate standardized tests, (2) teacher-made objective tests, and (3) teacher-made performance tests. To measure changes in attitude one can (1) observe (particularly by using outside observers from other schools), (2) use questionnaires to be answered by pupils or parents, (3) use rating scales such as the Meaning of Words Inventory developed at the University of California based on Osgood's Semantic Differential, (4) use dropout counts, (5) use records of parent involvement in school sponsored projects, (6) use case studies, or (7) anecdotal records, (8) use attendance records, and (9) use records of participation in activity. Evaluative criteria dealing with interest might be (1) questionnaires, (2) attendance records, (3) case studies, (4) anecdotal records, (5) dropout counts, (6) records of parent involvement, (7) various tabulations such as number of books read per pupil, and (8) rating scales. To measure work habits one could use (1) observation, (2) anecdotal records, and (3) rating scales or checklists. To measure personal and social adaptability one could use (1) dropout information, (2) attendance records, (3) anecdotal records, (4) rating scales, (5) pupils' writings, (6) sociograms, and (7) case studies. Other criteria are also useful such as those presented by Metfessel of USC at the 1963 AERA Convention, or the four point evaluation decision model.

One measurement that should not be overlooked is that of parent and teacher attitudes. Opinions and attitudes of parents are very important. Participation in conferences, attendance at school activities and other examples of adult behavior such as monitoring certain television programs, providing a quiet place for homework and use of library facilities provide indexes of opinion through choices of behavior. Professional judgments of teachers, specialists and supervisors can be obtained about children in a project with specially constructed rating scales. While the emphasis of the evaluation plan should be on the discovery of what happens to pupils, the effects of teacher attitude, behavior, and method with children in projects is also important. Such results may be easily observable before significant changes have taken place in pupils' educational attainment. Other examples of such kinds of effects are changed pupil attitudes and improved health.

An important consideration that should be kept in mind in relation to the evaluation section of the proposal is that it is a separate entity. The evaluation section of a proposal should be able to stand alone and should have its own objectives directly related to the objectives of the proposal. It should have its own methodology such as the Table 2 description of how to get to the data or where the data is available. It should have a concise methodology for analyzing and interpreting the data in terms of the objectives.

The Selection of Evaluation Measures

This section will deal with reliability, validity, cost, utility, reasonableness, and acceptance of tests by the school. These are taken into consideration when one selects tests for evaluation. One important fact shouldn't be overlooked --- when a testing company such as Educational Testing Service sets out to design a test, they don't sit behind their desks and write up items and try them out, they hire teachers who have probably written 50 or 100 tests in their lifetime and say, "Tell us, what would you ask if you wanted to measure this particular criteria?" Some of the best tests we have today are basically teacher-made tests. The role of the teacher in designing evaluation tests should never be discounted. Teachers know what the learning is in the classroom and if one wants to get an accurate picture of how much has been learned, for comparison, he can get it through no other means directly related to a program as well as through teacher-made tests.

There are certain other specific standardized tests which are especially useful in evaluation: (1) Special diagnostic tests - The Frostig Developmental Test of Visual Perception; Guilford's Group Test of Creativity; Maturity Levels for School Readiness and Reading Readiness; How-to-Study tests dealing with work habits; The Brenner Developmental Test of School Readiness; Diagnostic Reading tests such as Durrell and Gray, The Sequential Test of Educational Progress; The Iowa Test of Educational Development; the Evaluation and Adjustment Series; the California Reading Test; Auditory Discrimination Test by Wepman; the Bender-Gestalt. (2) Interest Tests - The Kuder Vocational; Edwards Personal Preference Schedule; behavior preference records set up by the students; the Strong Vocational Interest Blank for Men and Women. (3) Personality Tests - The Vineland Social Maturity Scale; SRA Youth Inventory; Mooney Problems Checklist; MMPI (Minnesota Multiphasic Inventory); California Test of Personality: Meaning of Words Inventory based on Osgood's Semantic Differential; Inventory of Self-appraisal; Minnesota Counseling Inventory; Short Form California Test of Mental Maturity; the regular California Test of Mental Maturity; Allport's Study of Values; Kuder Personal Records; Guilford-Zimmerman Temperament Schedule. (4) Achievement Tests - California Achievement Test; California Test in Social and Related Sciences; Essential High School Content Battery; Iowa Test of Basic Skills; Metropolitan Achievement Tests; Stanford Achievement Test; Wide Range Achievement Test; Metropolitan Language Test for Elementary Grades, SRA Achievement Test. (5) Personality or Projective Tests - The Rorschach; Sacks Sentence Completion Test; Rotter Incomplete Sentences; Madelaine Thomas Completion Stories Perry Point Scale; Draw-A-Person Test, Thematic Apperception Test. (6) I.Q. Tests - The Wechsler Intelligence Scale for Children; the Wechsler Adult Intelligence Scale, the Goodenough I.Q. Test; the Otis Short Form Test of Mental Maturity; Ohio State Psychological Examination; Kuhlmann-Anderson I.Q. Test, Lois Andyk I.Q. Test; Peabody Picture Vocabulary Test; the Kuhlmann-Finch I.Q. Test; the School and College Abilities Test. (7) Aptitude Tests - the Differential Aptitude Test; the School and College Abilities Test. (8) Special Purpose Tests - The Chicago Non-Verbal Examination; Wonderlic Personnel Tests, the Army General Classification Test; Tyler Study Skills Inventory; acary Test of Mechanical Ability; Halstead-Wepman Aphasia Screening Test; Full Range Picture Vocabulary Test; Grace Arthur Point Scale; California Algebra Aptitude Test.

As one can see, there are a number of evaluative measurements in many different areas with which one can evaluate school projects. The results, their interpretation, and the design of test data gathering make up the body of the evaluation.

Section Four

STATISTICAL REFRESHER

Definition of Measurement Terms

1. Arithmetic Mean - The sum of an array of scores divided by the number of scores.

2. Battery - A group of several tests of which the results are of value individually, in combination, and/or totally. Generally the inter-correlations of the tests are extremely low whereas the reliability co-efficients are extremely high.

3. Centile - A value on the scoring scale below which lie any given percentage of cases. Many people use the term percentile instead of the correct term, centile, although both mean the same thing.

4. Chi-Square - or X^2 - A means of estimating whether a given distribution differs from expected values to such a degree as to be evidence for the operation of non-chance factors. It is obtained by summing the quotients obtained by dividing the square of each difference between an actual and expected frequency by the expected frequency. The degrees of freedom for a chi-square are obtained by taking the number of rows in the table minus 1, times the number of columns in the table minus 1, not including the totals.

5. Correlation Coefficient (small r) - This is the most commonly used measure of relationship between paired facts or of the tendency of two or more variables or attributes to go together. It ranges from a -1 to a +1 through 0.0 which indicates no relationship. It is the measure of how two variables covary with one another. If the correlation is positive, strongly positive, when one variable goes up, so does the next. When strongly negative, when one variable goes up, the other one goes down.

6. Criterion - A standard that provides a basis for evaluating the validity of a test.

7. Cross-Validation - The process of checking whether a decision derived from one set of data is truly affective when this decision is applied to an independent but relevant data. It shouldn't be confused with cross-comparison which is the process of comparing the results from two different tests, with neither being considered the criterion instrument.

8. Culture-fair test - A test yielding results that are not culturally biased.

9. Culture-free test - A test yielding results that are not influenced by cultural background factors.

10. Diagnostic Test -A test intended for the separate measurement of a specific aspect of achievement in a single subject or field. They yield measures of specific skills, knowledges, or abilities underlying achievement within a broad subject. They are designed to identify particular strengths and weaknesses of an individual.

11. The discriminating power of a test - The ability of the test item to differentiate between individuals possessing much of the same characteristic from those possessing little of the characteristic.

12. Evaluation - The total broad changes in relation to major objectives of an educational program. Evaluation and measurement are not synonymous terms.

13. Error Variance - The portion of the variance of test scores that is related to the test's unreliability.

14. Educational loading - Weighing of a test content with factors specifically related to formal education.

15. Face Validity - Refers to the acceptability of the test and test situation by the examiner or user. In terms of apparent uses for which the test is to be put, another word for practicality. A test also has face validity when it appears to measure the variable to be tested.

16. Factor Analysis - A method of analyzing the intercorrelations among a set of variables such as test scores. It attempts to account for interrelationships in terms of underlying groupings based on the correlations, preferably fewer in number than the original variables, these combinations are then called "factors". It is principally a method for data reduction and reveals how much of the variation of each of the original measures arises from or is associated with each type of the hypothetical factors.

17. Frequency Distribution - The tabulation of scores from high to low, or low to high, showing the number of persons who obtain each score in a group of scores.

18. Grade Norm - The average test score obtained by pupils classified at a given grade placement.

19. Item Analysis - Any one of several methods used in test construction to determine how well a given test item discriminates among individuals different in some characteristic. The effectiveness of the test item depends upon 3 things: (1) the validity of the item in regard to curriculum content and educational objectives; that is, content validity, (2) the discriminating power of the item in regard to validity and its internal consistency, and (3) the difficulty of the item, usually established by phi coefficients of correlation.

20. Mean - Sum of the set of scores divided by the number of scores.

21. Measurement - The emphasis in measurement is upon single aspects of subject matter achievement or specific skills and abilities. Measurement and evaluation are not synonymous terms. The emphasis in evaluation is upon broad changes and major objectives of the educational program.

22. Median - The middle score in a set of ranked scores. It is the point above or below which an equal number of rank scores lie. It corresponds to the 50th percentile.

23. Mode - The score or value that occurs most frequently in a distribution.

24. Normal Distribution Curve - A derived curve based on the assumption that variations from the mean are by chance. It is bell-shaped in form and adopted as true because of its repeated recurrence in the frequency distributions of sets of measurements of human characteristics in psychology and education. It has many useful mathematical properties. In a normal distribution curve, scores are distributed symmetrically about the mean -- as many cases at various equal distances above the mean as below the mean -- and with cases concentrated near the average and decreasing in frequency the further one departs from it.

25. Norms - Summarized statistics that depict the test performance of a specific group. Grade, age, and percentile are the most common types of norms.

26. Percentile - One of the 99 point scores that divide a rank distribution into groups, each of which contains 1/100 of the scores. It is a point in a distribution below which falls the percent of cases indicated by the given percentile; thus the 73rd percentile denotes the score or point below which 73 per cent of the scores fall in this particular distribution of scores.

27. Power Test - A test which is designed to sample the range of an examinee's capacity in particular skills or abilities and which places minimal emphasis on time limits. A test in which a subject may take as long as he wishes and go into as much depth as he wishes.

28. Random Sample - A sample drawn in such a way that every member of the population has an equal chance of being included thus eliminating any selection bias. A random sample is generally thought of as being "representative" of its total population.

29. Range - The difference reflected by noting the lowest and highest scores obtained on a test by some group.

30. Reliability - The degree to which a pupil would obtain the same score if the test were readministered to the pupil (assuming no additional learning affects, etc.). The trustworthiness of scores. There are several types of reliability coefficients that should be distinguished. Type A - the coefficient of internal consistency, refers to a measure being based on internal analysis of data obtained on a single trial of the test. The more prominent of this method is the Kuder-Richardson-Hoyt Analysis of Variance and the split-half method. Type B - the coefficient of equivalence, refers to a correlation between scores from two forms of a test or parallel forms of the test that are essentially the same. Type C - the coefficient of stability, refers to a correlation between a test and retest with some period of time intervening. The test-retest situation may be with two forms of the same test. The coefficient of stability is the correlate of the first administration with the second administration.

31. Speed Test - A test in which performance is measured by the number of items performed in a given time. It is the opposite of a power test where time is of no importance.

32. Standard Deviation - This is a statistic used to express the extent of the deviations from the mean for the distributions. If the group tested is normal, their scores, when plotted, would yield a normal distribution curve. Two-thirds, or 68.3% of the scores would lie within the limits of one standard deviation above and one standard deviation below the mean. One-third of the scores would be above the mean by 1 standard deviation and one-third below the mean by 1 standard deviation. About 95% of the scores would lie within the limits of 2 standard deviations above and below the mean. About 99.7% of the cases would lie within the limits of 3 standard deviations above and below the mean.

33. Standardized Tests - A test that is composed of empirically selected materials that has definite directions for administration, scoring and use, data on reliability and validity, and adequately determined norms.

34. Standard score (commonly known as sigma score, t score or z score) - A score expressed as a deviation from the mean in terms of the standard deviation of the distribution. It is the raw score minus the mean divided by the standard deviation.

35. Stanine - A unit that divides the norm population in the normal distribution into 9 groups. Except for stanines 1 and 9, the groups are spaced in half standard deviation units with the mean at 5.

36. Stratified Sample - A sample in which cases are selected by the use of certain controls such as geographical regions, community size, grade, age, sex, etc.

37. Survey Test - A test that measures general achievement in a given subject area. It is used to test skills and ability of widely varying types. A survey test may also yield diagnostic information.

38. T Score - A derived score based on the equivalence of percentile values to standard scores; thus avoiding non-normal distributions. Usually has a mean equated to 50 and a standard deviation equal to 10.

39. Validity - The extent to which a test measures the trait for which it was designed or for which it is being used rather than some other trait.

    There are two basic approaches to the determination of validity. One is logical and one is empirical. Under logical validity, we have content validity which refers to how well the content of the test samples the subject matter or situation about which conclusions are to be drawn - characteristics of achievement tests primarily. Item structure is another form of logical approach to the determination of validity and this includes (1) corroborative evidence from item analysis supporting the other characteristics of the test; for example, intercorrelationships between items and items and scores, etc., and (2) item composition. Another approach to the determination of validity is empirical validity. There are two types of empirical validity: predictive and concurrent validity. Predictive validity relates to how well predictions from the test are confirmed by data collected at a later time; for example, predicting who will be good in medical school. Concurrent validity refers to how well test scores match measures of contemporary criterion performance; for example, comparing the distributions of scores for men in an occupation for those of men in general.

    There is a third type of validity which is both logical and empirical and referred to as "construct validity". It deals with the psychological qualities of test measures such as the personality measurement on a Rorschach Ink Blot Test. The Guilford Tests of Creativity are based on a construct - Guilford's factor structure of intelligence - so they have construct validity. Any time factor analysis is used, one has construct validity. Use of a personality or interest inventory to describe a person has construct validity.

40. Variability - The spread or dispersion of scores usually indicated by quartile deviations, standard deviations, range of 90 to 10 percentile scores, etc.

Statistical Introduction

Now that some of the terminology essential to the use of statistics has been defined, a refresher course on statistics and statistical analysis is helpful. In developing evaluation methodology, the research worker basically uses two types of statistics: descriptive, and inferential, or sampling. Descriptive statistics are used to do exactly what they say they are going to do, describe. Average, for example, gives information about the amount of certain qualities present in a group of individuals. It gives a basis for making comparisons between groups. Average is generally referred to as central tendency or central value. Two measures of central tendency are the mean and the median. Measures of dispersion indicate variability or scatter within a group of scores. One measure of dispersion is standard deviation. Another is variance. They answer the question as to what extent the scores tend to spread out within the distribution or to what extent they tend to cluster around the mean.

Inferential statistics have a number of important concepts. Among them are population, sample, standard error, level of significance and probablity. A population is a well-defined group of individuals or observations. A sample is a limited number of individuals from that defined population. A sample may be drawn randomly or it may be stratified by sex, religion, or ethnic group, so that the percentages are taken from the sample equal to the percentages in the population. Naturally stratified sampling is much more difficult to achieve than random sampling. It is for this reason that random sampling is more commonly employed in educational research. In school projects we no longer have a random sample, we have a stratified sample because one can't randomly select students and put them in projects. They are there. However, one can select randomly from the students already in the projects and by this means parcel out some of the error variance. One other alternative is to test all of the students. Inferential statistics allow the investigator to make inferences about a population based on data from a given sample. Sampling statistics indicate how well the statistics from a given sample probably represent the larger populations from which the samples were drawn.

In interpreting the results of research studies, level of significance becomes important. It is the most important aspect of statistical analysis. In applying a statistical test such as chi-square to the data from an investigation we are often interested in determining the degree to which chance factors explain the observed results. Probability is usually expressed in terms of the number of chances out of 100 that the observed results could be attributed to chance factors, expressed by a decimal fraction. A probability of .05 means five chances out of 100 and .01 probability would mean one chance out of 100 that you would receive a certain result by chance. Naturally .01 probability gives data more significance.

A significance level is actually a probability statement expressed as a per cent. If the probability of the given set of observation occurring by chances at the .05 (or 5 out of 100), we say that it is also significant at the 5 per cent level. For example, if we give a pre-test and a post-test of reading at the beginning of a project and at the end of a project, and compare the means of the pre- and the post-test to see if there are significant changes, or significant developments, from the beginning to the end of the project, we would run what is known as a t-ratio, or t statistic with a certain number of degrees of freedom which is based on the size of the sample at the beginning and at the end.

After we calculate this t-ratio, we would look in a table of t-values to find out whether the critical t-ratio was larger than the number needed to be significant at the .05 or the .01 level. If it is significant at the .01 level (which incidentally should be set before the research begins), we would then say there has been a significant difference from the beginning to the end. This is an inferential hypothesis. We are inferring that there is significant difference from the beginning to the end.

While the typical researcher ordinarily requires difference of significance at either the 5 per cent level or the 1 per cent level before rejecting the hypothesis of no significant differences between means, standard deviations, or other statistics, usually in a pilot study the .05 level is acceptable in order that meaningful results will not be screened out because of the lack of preciseness in the measuring instruments. In the final study, however, it is generally more acceptable to use the .01 level of significance since at that time sampling and instruments have become much more refined.

Once a sample has been gathered, and probability statements, and levels of significance set and data gathered, the researcher has two types of statistical methods at his disposal - parametric statistical methods and non-parametric statistical methods. The big difference is the normal Gaussian curve. Parametric statistics are based on certain known characteristics of this normal curve. The use of non-parametric statistics does not assume that the groups under study are random samples from a normally distributed population. As this assumption quite often cannot be met, these statistics are quite useful. When this is the case, one uses these non-parametric statistics which do not depend on the normal curve for their validity.

Probably the simplest test for the significance of the difference between two means is the t test, which is a parametric test. The t test requires that the distribution of scores is normal and that the test measured gives a continuous score. While theoretically the continuous score assumption is that the score could vary all the way from zero to infinity, in reality it is improbable that one would score infinitively on any test.

The use of the t test can be in the difference between the mean on a pre-test and the mean on a post-test for the same group of children, as was explained earlier. Or, it can be the difference between means of two randomly assigned groups subjected to different treatment, as was mentioned earlier. If one cannot assume that the samples being studied are drawn from a normal population, or if sample sizes are very small where one would not get a normal distribution, one should use an appropriate non-parametric test. One such test for two independent sample conditions is the Mann-Whitney U Test. For reference of non-parametric test see Siegel, Non-parametric Statistics for the Behavioral Sciences, published by McGraw-Hill in 1956. The Mann-Whitney U Test is the non-parametric counterpart of the t test. They both answer the same question. The Mann-Whitney, however, doesn't require a normal distribution and can be used when one has very small samples.

Another familiar non-parametric test which is used a great deal is the "chi-square" test which may be used to test for the significance of difference between independent groups. This test is probably most appropriately used when dealing with frequencies or head counts rather than the test scores. What is tested in chi-square is whether the frequencies observed in several categories (for example, the responses to an item by different people) are different from the frequencies expected on the basis of some hypothesis. Chi-square allows a test of the significance of the difference between the number who actually fit the category and the number expected to be in the category; for example, teachers who said they like teaching and wouldn't trade it, versus teachers who said they didn't like teaching; and principals, asking the same questions. These inferential statistics are all univariate models. In addition to the univariate models there is the multivariate extension which allows for the variance of all variables to be included in tests of significance. As these methods require extensive knowledge of matrix algebra, they will not be considered here. Anderson, An Introduction to Multi-Variate Statistics, is a useful reference.

Summary

The methods mentioned herein are only some of the minor tests available for evaluating the effects of school programs. Methodology involving one-way analysis of variance is also useful when one has a number of subjects being exposed to more than two different kinds of treatment. In this case, it is more efficient to use an analysis of variance than to compute a t test between all the means and all the groups. In another situation, one might have two different kinds of treatment. In this case also, it is more efficient to use an analysis of variance than to compute a t test between all the means and all the groups. In another situation, one might have two different groups of children being exposed to three different kinds of teaching instruction. This would be a two-way analysis of variance problem. There are such things as three-way analysis of variance problems which are not a primary concern at this writing.

It is hoped that these sections dealing with the evaluation of school projects will help some individuals to become further acquainted with one method for designing evaluations. Of course, it is not expected that through the use of this type of brief description every person would become a statistician; it is hoped, however, that this meets a need among our staff to understand more fully the methodology behind research and evaluation design.

JDN:ss
#37500
Rev. 7/3/69